Analyzing the NYC Subway Dataset

Section 1. Statistical Test

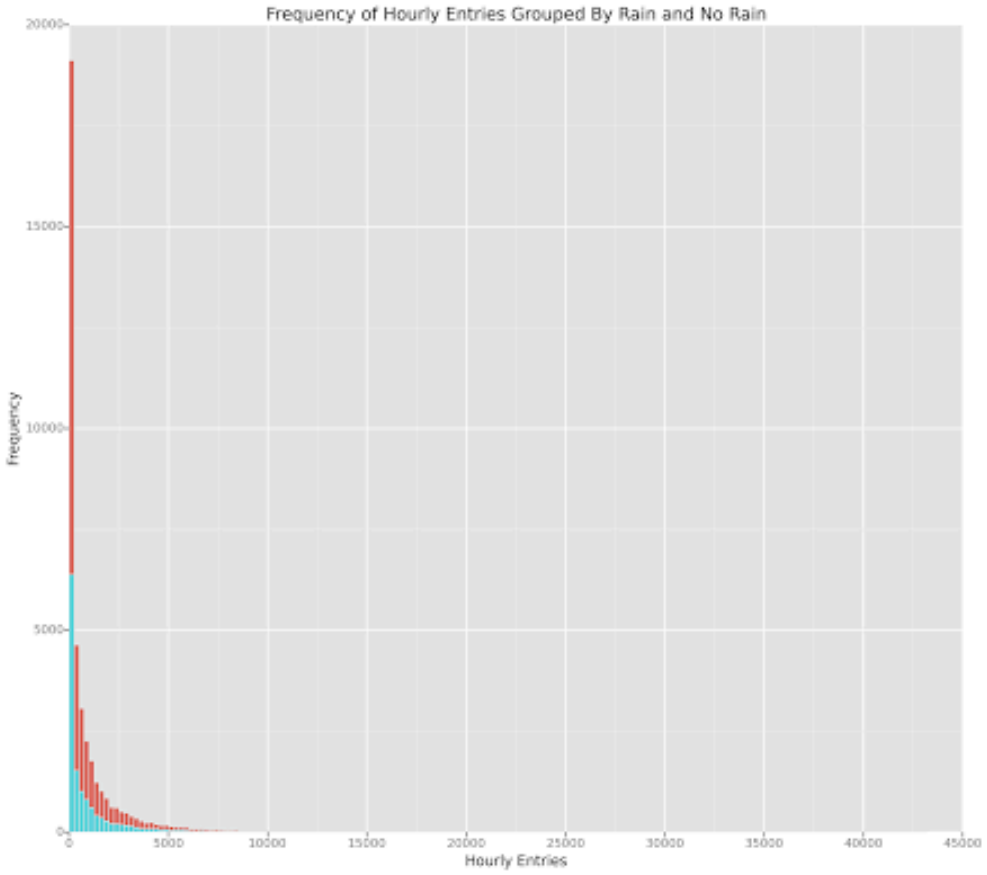
1. I used the Mann-Whitney U-test to analyze the NYC subway data and a two-tailed p-test. The null hypothesis for the Mann-Whitney U-test is that the hourly entries with rain are not significantly different from the hourly entries without rain. The p-critical value is 0.05.
2. The Mann-Whitney U-test is applicable to the dataset because the distribution of ridership is not a normalized distribution (it does not look like a Gaussian bell curve when plotted as a histogram). The Mann-Whitney U-test assumes non-normal distribution.
3. The results obtained from the Mann-Whitney U-test are as follows: the mean of the entries with rain was 1105.4463767458733, the mean of the entries without rain was 1090.278780151855, and the p-value was 0.024999912793489721.
4. These results mean that the distribution of the entries with rain is significantly different from the distribution of the entries without rain and that the null hypothesis must be rejected.

Section 2. Linear Regression

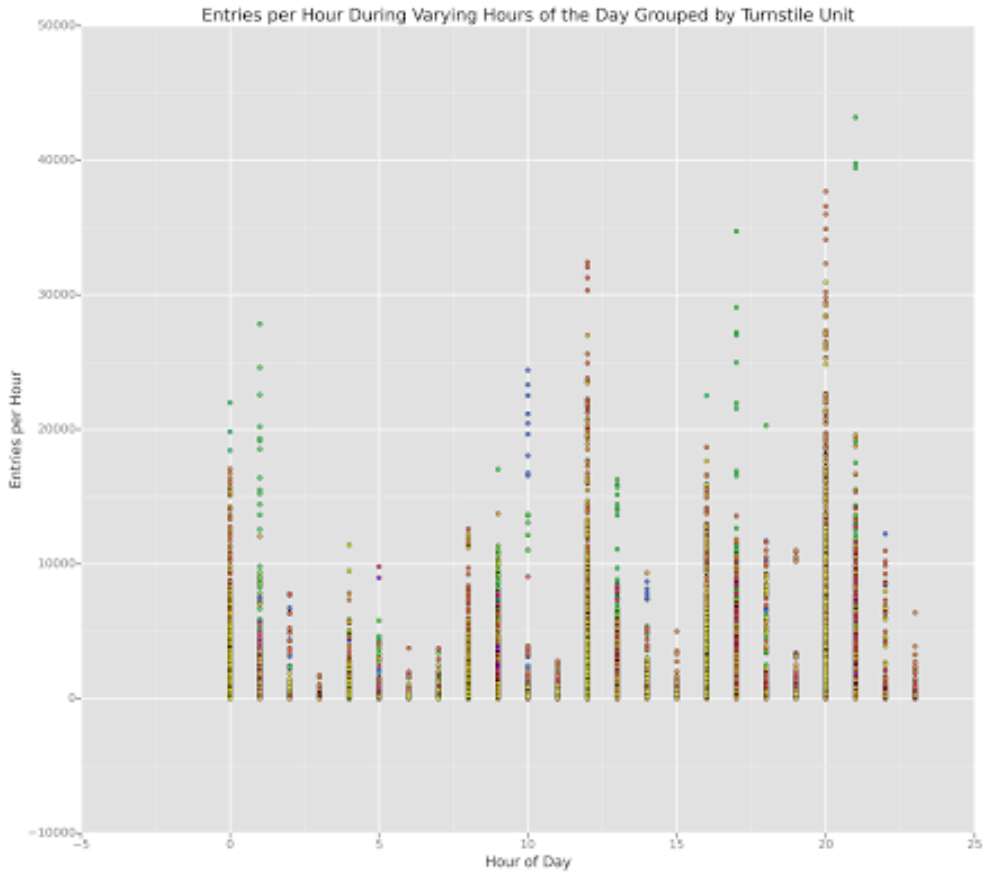
1. I used a) gradient descent to compute the coefficients theta and produce predictions for ENTRIESn_hourly in my regression model.
2. I used the rain, precipi, hour, meantempi and fog features in my model. I also used one dummy variable, that of UNIT.
3. The reason I used the first three features was because I thought that they were the ones that would have the most predictive ability. I thought that rain and precipi would predict to some degree that more people would ride the subway, and that hour would have strong predictive value because most people have set schedules and commutes. I added meantempi and fog once I looked at the coefficients they were weighted with. I think the day of the week would be another feature that would increase the predictive ability of this model, for a similar reason as the hour. I used the UNIT feature because it was already written in, and it made a huge difference.
4. The coefficients of my non-dummy features rain, precipi, hour, meantempi, and fog are -0.71229918, -0.40121265, -1.00510419, 1.47827891, and 2.24832915 respectively.
5. My model's R^2 value is 0.464492360273.
6. This R^2 value means that my regression model accounts for slightly more than 46.4% of the variance in the hourly entries of the subway. I think that this linear model to predict ridership is appropriate for this dataset given my R^2 value because it is predicting human behavior. People can be late and show up at a different hour than they usually do, they can be sick or move out of the city or drive or walk for a myriad of reasons. To account for almost

half of the variability given all of the different reasons people could have to ride or not ride the subway is appropriate.

Section 3. Visualization



1. This is a plot of two overlapping histograms, one of the frequency of increasing groups of hourly entries for when it wasn't raining, and one histogram of the same for when it was raining. I found this graph interesting because it shows that the distribution of the frequency for different sets of hourly entries for rain and no rain appear to be very close to one another.



2. This is a plot of the ridership by hour grouped by turnstile unit. This graph is interesting because the lines that stand out to me at first are multiples of four. These show similar color distributions and seem to be the 'main lines'. Following each of these, there is a general trend of less people riding the subway each hour until the next multiple of four. Also, the color distribution of the 1 + 4-hour group of units (hours 1, 5, 9, 13, 17, and 21) have a very different color distribution. At first I thought that the descending within the groups of four might imply that people were late, and less people were late 2 hours than were 1 hour and so on. But with the different color scheme this becomes less likely, because it represents a whole different group of subway stations with high hourly entries.

Section 4. Conclusion

1. From my analysis, there are multiple answers to the question "do more people ride the NYC subway when it is raining or when it is not raining?" In terms of absolute magnitude, more people ride the subway when it is not raining than when it is. This result is reflected in the first graph in section 3. The confounding factor in this is that it is more often not raining in NYC than it is raining. In problem set 3 we

looked at the average number of hourly entries when it was raining and when it wasn't. The results of this test showed that there was a significant difference between the two, and that the average hourly entries when it was raining were greater than the average hourly entries when it wasn't raining. Thus, in this sense, more people ride the NYC subway when it is raining.

Section 5. Reflection

1. One potential shortcoming concerning the dataset is I was limited to using only a portion of the dataset for many of the analyses. The bigger the sample size that I could use, the more confident I can be in the results. One potential shortcoming of the analysis is that for the gradient descent I reached a local minima rather than a global one. As for the linear regression, of course the biggest shortcoming is the assumption that the data will correlate in a linear fashion. This is not always the case, and is certainly a questionable assumption given the complexity of our data set.

References:

https://en.wikipedia.org/wiki/Normalization_%28statistics%29

https://docs.python.org/2/library/datetime.html

https://pypi.python.org/pypi/ggplot/

http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Mapper

http://www.dotnetperls.com/strip

http://www.tutorialspoint.com/python/string_split.htm

https://stackoverflow.com/questions/3691975/s-format-vs-0-format-vs-format

https://docs.python.org/2/library/logging.html#logger-objects

https://docs.python.org/2.6/library/logging.html#logging.info

http://www.tutorialspoint.com/python/python_loop_control.htm

https://stackoverflow.com/questions/2025701/conditional-skip-in-python-if

http://zetcode.com/lang/python/itergener

https://stackoverflow.com/questions/1733004/python-next-function

https://docs.python.org/3/library/functions.html#next

http://www.tutorialspoint.com/python/membership_operators_example.htm

https://stackoverflow.com/questions/21086736/sys-stdin-readline-reads-without-prompt-returning-nothing-in-between

https://docs.python.org/2/library/sys.html

https://stackoverflow.com/questions/1602934/check-if-a-given-key-already-exists-in-a-dictionary

https://docs.python.org/2/tutorial/datastructures.html#dictionaries

https://stackoverflow.com/questions/1024847/add-key-to-a-dictionary-in-python

http://pymotw.com/2/logging/

https://stackoverflow.com/questions/265960/best-way-to-strip-punctuation-from-a-string-in-python

http://www.cookbook-r.com/Graphs/Titles_(ggplot2)/

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.join.html

https://en.wikipedia.org/wiki/Coefficient_of_determination

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

https://bespokeblog.wordpress.com/2011/07/11/basic-data-plotting-with-matplotlib-part-3-histograms/

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.hist.html

https://stackoverflow.com/questions/6986986/bin-size-in-matplotlib-histogram

http://docs.scipy.org/doc/numpy/reference/generated/numpy.histogram.html

https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

http://pandas.pydata.org/pandas-docs/stable/comparison_with_sql.html

http://matplotlib.org/api/pyplot_api.html

http://ggplot.yhathq.com/how-it-works.html

http://ggplot.yhathq.com/docs/geom_histogram.html

http://ggplot.yhathq.com/docs/index.html

http://docs.ggplot2.org/current/geom_histogram.html

http://docs.ggplot2.org/current/geom_bar.html

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html

https://stackoverflow.com/questions/6919025/how-to-assign-colors-to-categorical-variables-in-ggplot2-that-have-stable-mappin

https://stackoverflow.com/questions/25604115/dataframe-constructor-not-properly-called-error

http://pandas.pydata.org/pandas-docs/dev/dsintro.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.dot.html

https://stackoverflow.com/questions/354883/how-do-you-return-multiple-values-in-python

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#scipy.stats.ttest_ind

http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.chisquare.html

http://docs.scipy.org/doc/scipy/reference/stats.html

http://pandas.pydata.org/pandas-docs/version/0.15.2/10min.html

https://www.mathsisfun.com/data/standard-normal-distribution.html

https://stackoverflow.com/questions/22950704/pandas-how-to-write-csv-file-using-data-from-a-dataframe-as-part-of-path-filena

https://docs.python.org/2/library/string.html

https://en.wikipedia.org/wiki/Simpson%27s_paradox

http://blog.mikiobraun.de/2013/11/how-python-became-the-language-of-choice-for-data-science.html

https://stackoverflow.com/questions/993984/why-numpy-instead-of-python-lists

https://stackoverflow.com/questions/19798153/difference-between-map-applymap-and-apply-methods-in-pandas

https://stackoverflow.com/questions/26394363/dot-product-between-a-matrix-and-a-1d-array-in-numpy?rq=1

https://stackoverflow.com/questions/11033573/difference-between-numpy-dot-and-inner

https://stackoverflow.com/questions/3023649/hour-from-datetime-in-24-hours-format

http://www.tldp.org/LDP/Bash-Beginners-Guide/html/Bash-Beginners-Guide.html

https://stackoverflow.com/questions/2429511/why-do-people-write-usr-bin-env-python-on-the-first-line-of-a-python-script

https://docs.python.org/2/library/csv.html

http://www.w3schools.com/sql/sql_groupby.asp

http://www.w3schools.com/sql/sql_func_count.asp

https://dev.mysql.com/doc/refman/5.1/en/counting-rows.html